Unifying Data Units and Models in Statistics
Focus on (Co-)Clustering

C. Biernacki
(with A. Lourme)

Workshop on Model-Based Clustering and Classification
September 5-7, 2016, Catania (Italy)

Take home message

Units are entirely interrelated with models

This talk:

- Be aware that interpretation of ("classical") models is unit dependent
- Models should even be revisited as a couple units $\times$ "classical" models
- Opportunity for cheap/wide/meaningful enlarging of "classical" model families
- Focus on model-based (co-)clustering but larger potential impact

Outline

General (model-based) statistical framework

- Data:
  - Let $\mathbf{d} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n)$, with $\boldsymbol{d}_i \in \mathcal{D}$
  - Each $\boldsymbol{d}_i$ value is provided with a unit **id**
  - We note "**id**" since units are often user defined (a kind of canonical units)
- Statistical aim: estimate a target quantity $\mathbf{t}$
- Model-based resolution: use a pdf[1] family $\mathrm{p}^{\mathbf{m}}$ associated to a model $\mathbf{m}$
- Final estimate of $\mathbf{t}$:

$$\hat{\mathbf{t}} = \hat{\mathbf{t}}(\mathbf{d}, \mathbf{m})$$

- Evaluate the model $\mathbf{m}$: use an information criterion $\mathbf{C}(\hat{\mathbf{t}})$

[1] probability density function

Changing the data units

- Principle of data units transformation $\mathbf{u}$:

$$\begin{array}{rccc} \mathbf{u}: & \mathbb{D} = \mathbb{D}^{\mathbf{id}} & \longrightarrow & \mathbb{D}^{\mathbf{u}} \\ & \mathbf{d} = \mathbf{d}^{\mathbf{id}} = \mathbf{id}(\mathbf{d}) & \longmapsto & \mathbf{d}^{\mathbf{u}} = \mathbf{u}(\mathbf{d}) \end{array}$$

- Question 1: has the unit change a consequence on the estimate target?

$$\hat{\mathbf{t}}(\mathbf{d}^{\mathbf{u}}, \mathbf{m}) \stackrel{?}{=} \hat{\mathbf{t}}(\mathbf{d}, \mathbf{m})$$

- Question 2: if yes, how to manage it for the statistical framework at hand?

Outline

Predictive target

Rewrite the general statistical case in this particular situation:

- Data:
  - Let $\mathbf{d} = (\mathbf{x}, \mathbf{y})$ with $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\mathbf{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ and $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbb{X} \times \mathbb{Y}$
  - Each $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ value is provided with a unit $\mathbf{id} = (\mathbf{id}_x, \mathbf{id}_y)$
- Statistical aim: estimate the predictive pdf $\mathbf{t} = \mathrm{p}(\boldsymbol{y}|\boldsymbol{x})$
- Model-based resolution: use a pdf family $\mathrm{p}^{\mathbf{m}}(\boldsymbol{y}|\boldsymbol{x})$ associated to a model $\mathbf{m}$
- Final estimate of $\mathbf{t}$:

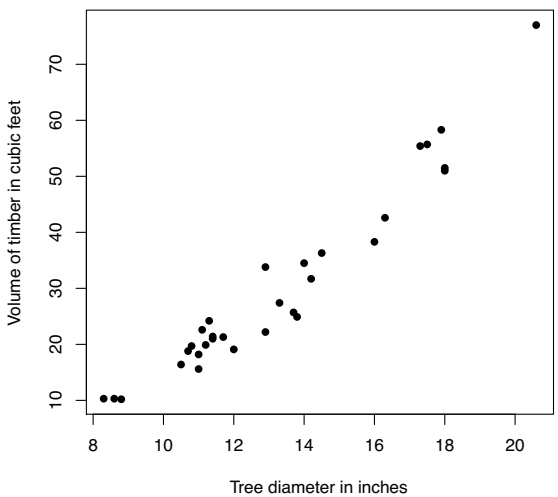
$$\hat{\mathbf{t}} = \hat{\mathbf{t}}(\mathbf{x}, \mathbf{y}, \mathbf{m})$$

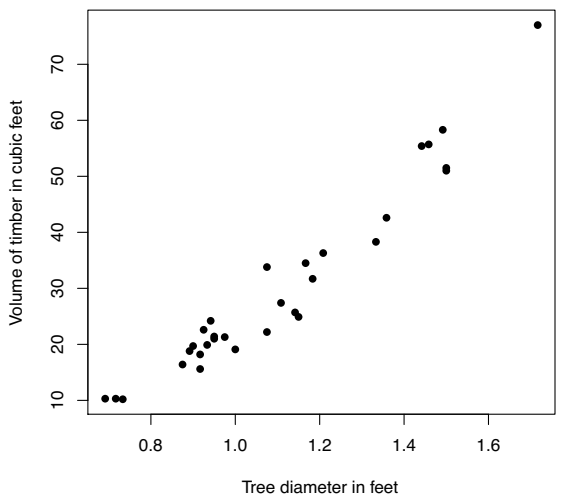
- Evaluate the model $\mathbf{m}$: use typically criteria $\mathbf{C} \in \{\mathrm{CV/PRESS}, \mathrm{AIC} \ldots\}$

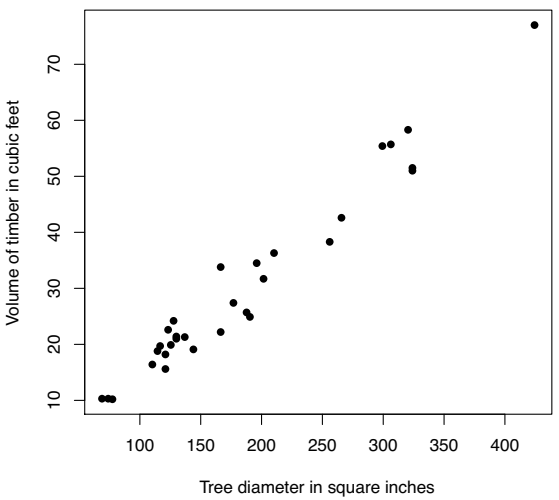
A linear regression example with three different units

- Data: measurements of $n = 31$ felled black cherry trees[2] where
  - x="girth" (tree diameter measured at 4 ft 6 in above the ground)
  - $\mathbf{id}_x$="inches" is the initial unit, $\mathbb{X} = \mathbb{R}^+$
  - y="volume of timber"
  - $\mathbf{id}_y$="cubic ft" is the initial unit, $\mathbb{Y} = \mathbb{R}^+$
- Three units $u_{x,j}$ on x ($j \in \{1, 2, 3\}$): note that $\mathbb{X}^{\mathbf{u}_{x,j}} = \mathbb{R}^+$ for all j

$\mathbf{u}_{x,1}$	$\mathbf{u}_{x,2}$	$\mathbf{u}_{x,3}$
"inches"	"feet"[3]	"square inches"
$\mathbf{id}_x(\cdot)$	$(\cdot/12)$	$(\cdot)^2$







[2] Atkinson, A. C. (1985) Plots, Transformations and Regression. Oxford University Press.

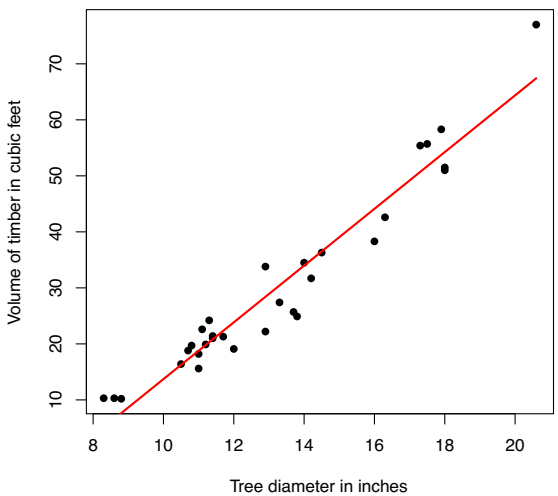
[3] 1 foot = 12 inches

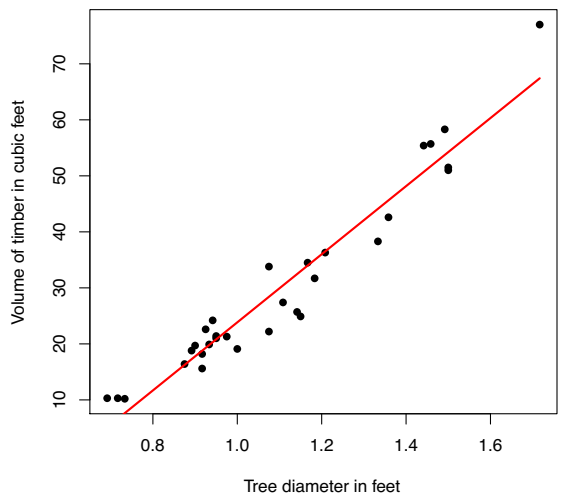
A linear regression for each new unit

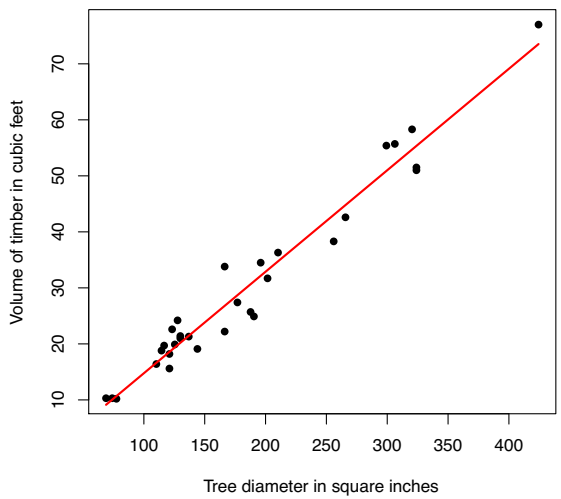
- Statistical aim: estimate the conditional probability $\mathrm{p}(y|x)$
- Model: Gaussian[4] linear regression with regards to each unit $\mathbf{u}_{x,j}$

$$\mathbf{m} = \{\phi(\cdot; \beta_0 + \beta_1 \underbrace{\mathbf{u}_{x,j}(x)}_{x\text{-axis unit}}, \sigma^2), \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}, \sigma^2 \in \mathbb{R}^{+*}\}$$

- Estimate: maximum likelihood (ml) $\hat{\mathbf{t}}(\mathbf{u}_{x,j}(\mathbf{x}), \mathbf{y}, \mathbf{m})$







[4] $\phi(\cdot; \mu, \sigma^2)$ is the Gaussian density of mean μ and variance σ^2

About question 1

Has the unit change a consequence on the estimate target?

$$\hat{\mathbf{t}}(\mathbf{u}_j(\mathbf{d}), \mathbf{m}) \stackrel{?}{=} \hat{\mathbf{t}}(\mathbf{d}, \mathbf{m})$$

$$\begin{aligned}
\hat{\mathbf{t}}(\mathbf{u}_{x,1}(\mathbf{x}), \mathbf{y}, \mathbf{m}) &= \phi(y; -36.94 + 5.06\ \mathbf{u}_{x,1}(x), 16.91) \\
&= \phi(y; -36.94 + 5.06\ x, 16.91) \\
\hat{\mathbf{t}}(\mathbf{u}_{x,2}(\mathbf{x}), \mathbf{y}, \mathbf{m}) &= \phi(y; -36.94 + 60.79\ \mathbf{u}_{x,2}(x), 16.91) \\
&= \phi(y; -36.94 + 60.79\ \left(\frac{x}{12}\right), 16.91) \\
&= \phi(y; -36.94 + 5.06\ x, 16.91) \\
\hat{\mathbf{t}}(\mathbf{u}_{x,3}(\mathbf{x}), \mathbf{y}, \mathbf{m}) &= \phi(y; -3.35 + 0.18\ \mathbf{u}_{x,3}(x), 10.62) \\
&= \phi(y; -3.35 + 0.18\ (x)^2, 10.62)
\end{aligned}$$

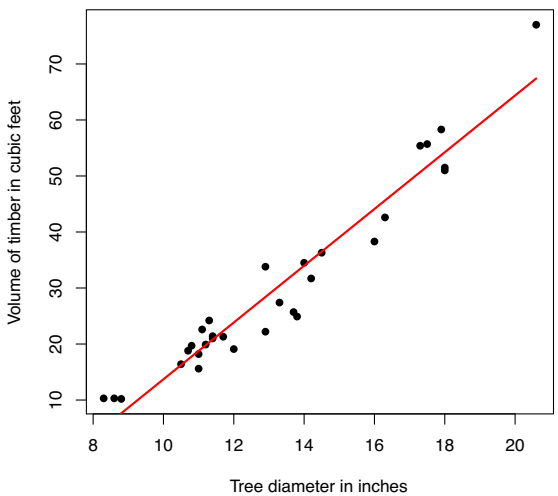
	$\hat{\mathbf{t}}(\mathbf{u}_{x,2}(\mathbf{x}), \mathbf{y}, \mathbf{m})$	$\hat{\mathbf{t}}(\mathbf{u}_{x,3}(\mathbf{x}), \mathbf{y}, \mathbf{m})$
$\hat{\mathbf{t}}(\mathbf{u}_{x,1}(\mathbf{x}), \mathbf{y}, \mathbf{m})$	equal	different

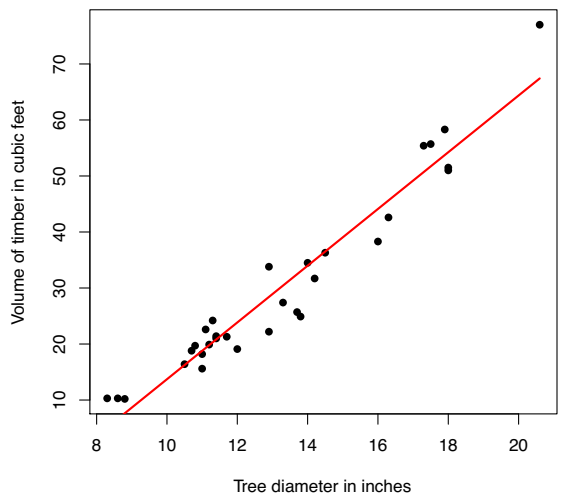
Possibly non-linear regression with the initial unit

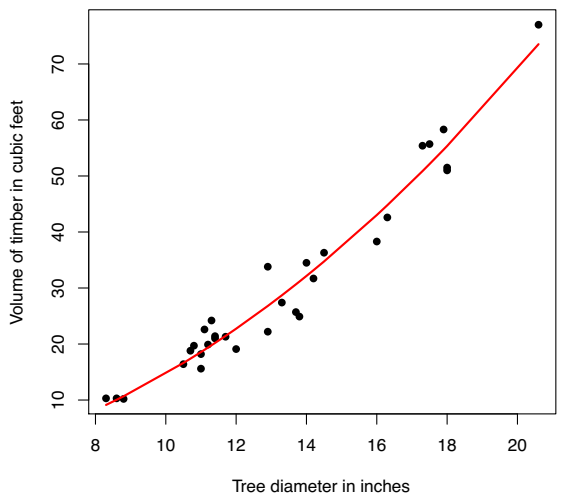
New unit $\Longleftrightarrow$ new model

	new unit		*initial unit*
$\mathbf{u}^{-1}$:	$\mathbb{D}^{\mathbf{u}_j}$	$\longrightarrow$	$\mathbb{D} = \mathbb{D}^{\mathbf{id}}$
	$\mathbf{d}^{\mathbf{u}_j} = \mathbf{u}_j(\mathbf{d}) = (\mathbf{id}_y, \mathbf{u}_{x,j}(x))$	$\longmapsto$	$\mathbf{d} = \mathbf{u}_j^{-1}(\mathbf{d}^{\mathbf{u}_j}) = (\mathbf{id}_y, \mathbf{u}_{x,j}^{-1}(u_{x,j}(x)))$
	$Y \vert \mathbf{u}_{x,j}(x) \sim \mathrm{p}^{\mathbf{m}}$		$Y \vert x \sim \text{“}\mathbf{u}_j^{-1}(\hat{\mathrm{p}}^{\mathbf{m}})\text{”} = \mathrm{p}^{\text{“}\mathbf{u}_j^{-1}(\mathbf{m})\text{”}} = \mathrm{p}^{\mathbf{m}_j}$
	initial (linear) model $\mathbf{m}$		new model “$\mathbf{m}_j = \mathbf{u}_j^{-1}(\mathbf{m})$”

$\mathbf{m}_1$	$\mathbf{m}_2$	$\mathbf{m}_3$
linear	linear	quadratic







About question 2

How to manage the consequence of the unit change?

Since "$\mathbf{m}_j = \mathbf{u}_j^{-1}(\mathbf{m})$" produces a new model $\mathbf{m}_j$

1. model design: create new models $\mathbf{m}_j$ by combining standard sets $\{\mathbf{u}_j\}$ and $\{\mathbf{m}\}$
2. model interpretation: $\mathbf{m}_j$ is meaningful if unit $\mathbf{u}_j$ and model $\mathbf{m}$ are meaningful
3. model selection: simply select $\mathbf{m}_j$ with any model selection criterion $\mathbf{C}$

Return to the example:

1. model design: "$\mathbf{m}_3 = (\text{square inches})^{-1}(\text{linear})$"
2. model interpretation: $\mathbf{m}_3$ decomposition is not unique!

$\mathbf{u}_{x,j} \backslash \mathbf{m}$	linear	quadratic
inches	$\{\mathbf{m}_1, \mathbf{m}_2\}$	$\mathbf{m}_3$
feet	$\{\mathbf{m}_1, \mathbf{m}_2\}$	$\mathbf{m}_3$
square inches	$\mathbf{m}_3$	new model $\mathbf{m}_4$!

3. model selection:

	$\mathbf{m}_1$	$\mathbf{m}_2$	$\mathbf{m}_3$
PRESS	637.52	637.52	379.58
AIC	181.64	181.64	167.22

$\mathbf{m}_3$ is preferred and it makes sense: volume is linearly linked to the surface

Outline

Clustering target

Rewrite the general statistical case in this particular situation:

- Data:
  - Let $\mathbf{d} = \mathbf{x}$ with $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\boldsymbol{x}_i \in \mathbb{X}$
  - Each $\boldsymbol{x}_i$ value is provided with a unit **id**
- Statistical aim: estimate the hidden partition in g classes $\mathbf{t} = \mathbf{z} = (z_1, \ldots, z_n)$, where $z_i \in \{1, \ldots, g\}$ indicates the class number
- Model-based resolution: use a mixture model $\mathbf{m}$ of parameter $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\alpha}_k\}_{k=1}^{g}$

$$\mathrm{p}^{\mathbf{m}}(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \mathrm{p}(\boldsymbol{x}; \boldsymbol{\alpha}_k)$$

  where $\pi_k = \mathrm{p}(Z = k)$ and $\mathrm{p}(\boldsymbol{x}; \boldsymbol{\alpha}_k) = \mathrm{p}(\boldsymbol{X} = \boldsymbol{x} | Z = k)$
- Final estimate of $\mathbf{t}$: from the ml estimate $\hat{\boldsymbol{\theta}}^{\mathbf{m}}$ (for instance)

$$\hat{\mathbf{t}} = \hat{\mathbf{z}}(\mathbf{x}, \mathbf{m}) \quad \text{where} \quad \hat{z}_i = \arg \max_{k \in \{1, \ldots, g\}} \mathrm{p}(Z_i = k | \boldsymbol{X}_i = \mathbf{x}_i; \hat{\boldsymbol{\theta}}^{\mathbf{m}})$$

- Evaluate the model $\mathbf{m}$: use typically criteria $\mathbf{C} \in \{\mathrm{BIC}, \mathrm{ICL} \ldots\}$

Recall the unit transformation principle

Combine $\mathbf{u}$ (bijective) and $\mathbf{u}^{-1}$, similarly to the previous predictive example

initial unit		*new unit*		*initial unit*
$\mathbb{X}$	$\xrightarrow{\mathbf{u}}$	$\mathbb{X}^{\mathbf{u}}$	$\xrightarrow{\mathbf{u}^{-1}}$	$\mathbb{X}$
$\mathbf{x}$	$\longmapsto$	$\mathbf{x}^{\mathbf{u}}$	$\longmapsto$	$\mathbf{x}$
–		$\mathbf{u}(\mathbf{X}) \sim \mathrm{p}^{\mathbf{m}}$		$\mathbf{X} \sim$ "$\mathbf{u}^{-1}(\mathrm{p}^{\mathbf{m}})$" $= \mathrm{p}^{\text{"}\mathbf{u}^{-1}(\mathbf{m})\text{"}} = \mathrm{p}^{\mathbf{m}'}$
no model		"standard" model $\mathbf{m}$		new model "$\mathbf{m}' = \mathbf{u}^{-1}(\mathbf{m})$"

We will discuss later also:

- Unit $\mathbf{u}$ can itself depend on a parameter $\boldsymbol{\lambda}$ to be estimated in the process ($\mathbf{u}_{\boldsymbol{\lambda}}$)
- Unit $\mathbf{u}$ can itself depend on the partition $\mathbf{z}$: $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_g)$
- Allowed units $\{\mathbf{u}\}$ depend on the data feature (continuous, binary, integer...)

Essential remarks

1 Model interpretation

- Two different interpretations of the previous transformation:
  - either as model $\mathbf{m}$ with units $\mathbf{u}$
  - or as model $\mathbf{m}' = \mathbf{u}^{-1}(\mathbf{m})$ with unit $\mathbf{id}$
- Consequences:
  - Always have a read to a model $\mathbf{m}$ with regards to its unit $\mathbf{u}$ (both are embedded)
  - Decomposition of $\mathbf{m}'$ into $\mathbf{u} \times \mathbf{m}$ can be more than two: choose the most meaningful!
  - Thus non identifiability of the decomposition...

2 Model design

- Conversely, opportunity to build easily numerous new meaningful models:
  - Just combine a standard model family $\{\mathbf{m}\}$ with a standard unit family $\{\mathbf{u}\}$
  - New family can be huge! Combinatorial problems can occur...
  - Some model stability can exist in some (specific) cases: $\mathbf{m} = \mathbf{u}^{-1}(\mathbf{m})$

3 Model selection

- Model selection with likelihood based criteria (BIC, ICL...):
  - Prohibited to compare $\mathbf{m}_1$ in unit $\mathbf{u}_1$ and $\mathbf{m}_2$ in unit $\mathbf{u}_2$
  - But allowed after transforming in identical unit $\mathbf{id}$: $\mathbf{m}_1' = \mathbf{u}_1^{-1}(\mathbf{m}_1)$ and $\mathbf{m}_2' = \mathbf{u}_2^{-1}(\mathbf{m}_2)$
  - Example for continuous $\boldsymbol{x}$ and differentiable $\mathbf{u}$: the density transform in $\mathbf{id}$ is the following

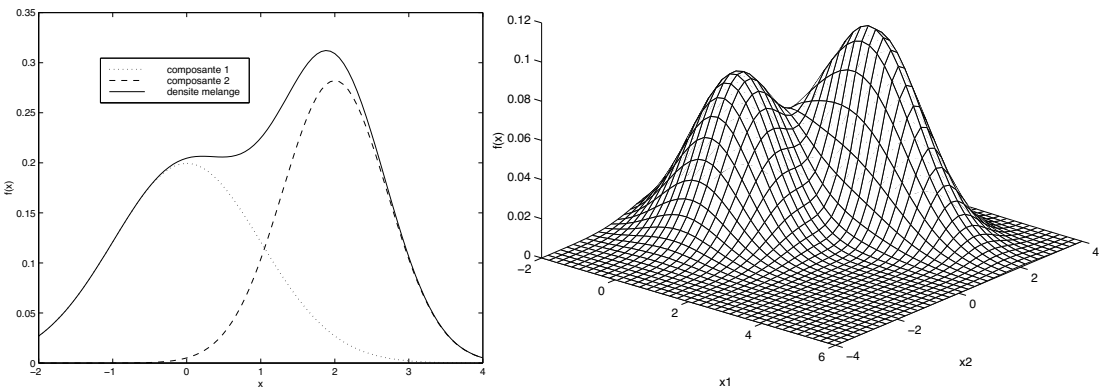
$$\mathrm{p}^{\mathbf{u}^{-1}(\mathbf{m})}(\boldsymbol{x};\boldsymbol{\theta}) = \mathrm{p}^{\mathbf{m}}(\boldsymbol{x}^{\mathbf{u}};\boldsymbol{\theta}) \times |\mathbf{J}^{\mathbf{u}}|$$

where $\mathbf{J}^{\mathbf{u}}$ is the Jacobian of the transformation $\mathbf{u}$

Outline

d-variate Gaussian mixtures

- $\boldsymbol{x} = (x^1, \ldots, x^d) \in \mathbb{X} = \mathbb{R}^d$
- d-variate Gaussian model $\mathbf{m}$: $\mathrm{p}(\cdot; \boldsymbol{\alpha}_k) = \mathcal{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$



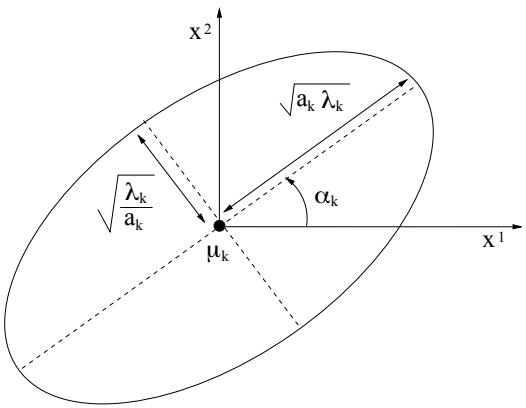
14 EIG models on $\mathbf{\Sigma}_k$

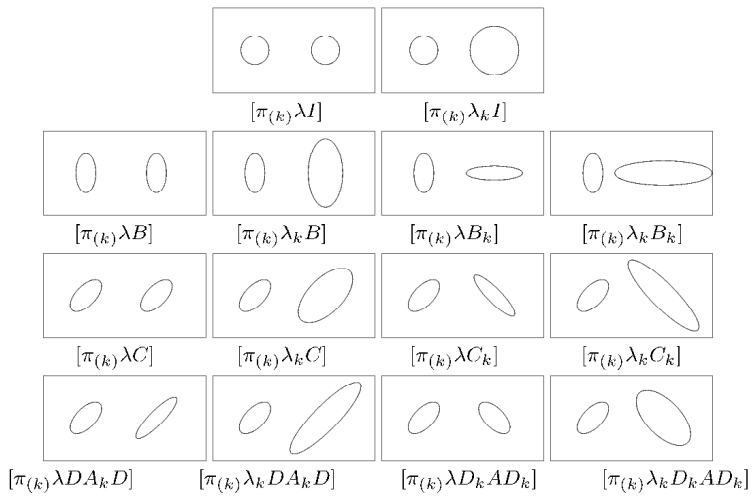
[Celeux & Govaert, 1995][5] propose the following eigen decomposition

$$\mathbf{\Sigma}_k = \underbrace{\lambda_k}_{\text{volume}} \cdot \underbrace{\mathbf{D}_k}_{\text{orientation}} \cdot \underbrace{\mathbf{\Lambda}_k}_{\text{shape}} \cdot \mathbf{D}_k'$$

where

- $\lambda_k = |\mathbf{\Sigma}_k|^{1/d}$
- $\mathbf{D}_k$ is an orthogonal matrix the columns of which are the eigenvectors of $\mathbf{\Sigma}_k$
- $\mathbf{\Lambda}_k$ is a diag. p.d. matrix with det. 1 and with diag. coef. in decreasing order





[5] Celeux, G., and Govaert, G.. Gaussian parsimonious clustering models. Pattern Recognition, 28(5), 781–793 (1995).

12 MFA models on $\boldsymbol{\Sigma}_k$

[Ghahramani & Hinton, 97][6], [McLachlan *et al.*, 03][7] propose the following mixture of factor analysers decomposition

$$\boldsymbol{\Sigma}_k = \mathbf{B}_k \mathbf{B}_k' + \omega_k \boldsymbol{\Lambda}_k$$

where

- $\mathbf{B}_k$ is a *loadings* $d \times q$ non-square real matrix ($1 \leq q \leq q_{\max}$, $q_{\max} < d$)
- ω_k is a positive real number
- $\boldsymbol{\Lambda}_k$ is a $d \times d$ diagonal positive definite matrix such that $|\boldsymbol{\Lambda}_k| = 1$

- Such models are essentially designed for high dimension thanks to their parsimony
- 12 parsimonious versions are then introduced by [McNicholas & Murphy, 10][8]:

$$[\mathbf{B}_k, \omega_k, \boldsymbol{\Lambda}]\{q\}, [\mathbf{B}, \omega, \boldsymbol{\Lambda}_k]\{q\}, [\mathbf{B}_k, \omega_k, \mathbf{I}]\{q\} \ldots$$

[6] Ghahramani, Z., Hinton, G.E. The EM algorithm for factor analyzers. Technical Report CRG-TR-96-1, University of Toronto (1997).

[7] McLachlan, G. and Peel, D. Modelling high-dimensional data by mixtures of factor analyzers. Computational Statistics & Data Analysis 41 (2003), 379–388.

[8] McNicholas, P.D., and Murphy, T.B. Model-based clustering of microarray expression data via latent Gaussian mixture models. Bioinformatics, 26(21), 2705–2712 (2010).

11 RTV models on Σ_k (and μ_k)

[Biernacki & Lourme, 2014][9] propose the following "statistical" decomposition

$$\Sigma_k = \mathbf{T}_k \mathbf{R}_k \mathbf{T}_k \quad \mu_k = \mathbf{T}_k \mathbf{V}_k$$

where (note: it is not Cholesky's decomposition)

- $\mathbf{T}_k$ is the corresponding diagonal matrix of conditional standard deviations
- $\mathbf{R}_k$ is the associated matrix of conditional correlations
- $\mathbf{V}_k$ gathers standardized means

- Statistical interpretation of the decomposition
- Possible to combine meaningful constraints on $\mathbf{T}_k$, $\mathbf{R}_k$ and μ_k (centers):
  - $\mathbf{T}_k$: free, isotropic ($\forall\ k : \mathbf{T}_k = a_k \mathbf{T}_1$ where $a_k > 0$[10]) or homogeneous ($\mathbf{T}_k = \mathbf{T}$)
  - $\mathbf{R}_k$: free or homogeneous ($\mathbf{R}_k = \mathbf{R}$)
  - Vectors $\mathbf{V}_k = \mathbf{T}_k^{-1} \mu_k$ ($k = 1, \ldots, K$): free or homogeneous ($\mathbf{V}_k = \mathbf{V}$)
- Notations: $[\mathbf{R}_k, \mathbf{T}_k, \mathbf{V}_k]$, $[\mathbf{R}, \mathbf{T}, \mathbf{V}_k]$, $[\mathbf{R}_k, a_k \mathbf{T}, \mathbf{V}_k]$...

[9] C. Biernacki and A. Lourme (2014). Gaussian Parsimonious Clustering Models Scale Invariant and Stable by Projection. Statistics and Computing, Volume 24, Issue 6, pp 953–969.

[10] Contrary to appearances, this model is invariant to the choice of the population numbering.

Scale unit invariance

- Consider scale unit transformation $\mathbf{u}(\mathbf{x}) = \mathbf{D}\mathbf{x}$, with diagonal $\mathbf{D} \in \mathbb{R}^{d \times d}$
- Very current transformation: standard units (mm, cm), standardized units
- [Biernacki & Lourme, 2014] listed models in each family where invariance holds

$$\mathbf{m} = \mathbf{u}^{-1}(\mathbf{m})$$

- Not invariant models produce new models
- Do not forget to compare all models $\mathbf{m}' = \mathbf{u}^{-1}(\mathbf{m})$ in unit **id** for BIC / ICL validity
- Used packages:
  - EIG: the `Rmixmod` R package
  - MFA: the `pgmm` R package
  - RTV: the `mixrtv` Matlab package

Family	invariant models
EIG	8 (among 14)
MFA	8 (among 12)
RTV	11 (among 11)

Illustration on the Old Faithful geyser data set

- All models are with free proportions (π_k)
- All ICL values are expressed with the initial unit min$\times$min
- We observe the effect of unit on the ICL ranking for EIG and MFA family
- Finally, it is the user responsability to choose between
  - opportunity to find new models by EIG and MFA (with better ICL value)
  - not new models by RTV but models benefiting from more invariance properties

family	rank	model	*ICL*	model	*ICL*	model	*ICL*
EIG	1	$[\lambda_k \boldsymbol{S} \boldsymbol{\Lambda}_k \boldsymbol{S}']$	1158.7	$[\lambda_k \boldsymbol{S} \boldsymbol{\Lambda}_k \boldsymbol{S}']$	1158.7	$[\lambda_k \boldsymbol{S}_k \boldsymbol{\Lambda} \boldsymbol{S}_k']$	1160.3
	2	$[\lambda_k \boldsymbol{S}_k \boldsymbol{\Lambda}_k \boldsymbol{S}_k']$	1161.4	$[\lambda_k \boldsymbol{S}_k \boldsymbol{\Lambda}_k \boldsymbol{S}_k']$	1161.4	$[\lambda_k \boldsymbol{S}_k \boldsymbol{\Lambda}_k \boldsymbol{S}_k']$	1161.4
	3	$[\lambda_k \boldsymbol{S} \boldsymbol{\Lambda} \boldsymbol{S}']$	1161.7	$[\lambda \boldsymbol{S} \boldsymbol{\Lambda}_k \boldsymbol{S}']$	1161.4	$[\lambda_k \boldsymbol{S} \boldsymbol{\Lambda} \boldsymbol{S}']$	1161.7
	4	$[\lambda_k \boldsymbol{S}_k \boldsymbol{\Lambda} \boldsymbol{S}_k']$	1160.3	$[\lambda_k \boldsymbol{S} \boldsymbol{\Lambda} \boldsymbol{S}']$	1161.7	$[\lambda \boldsymbol{S}_k \boldsymbol{\Lambda} \boldsymbol{S}_k']$	1162.8
MFA	1	$[\boldsymbol{B}, \omega_k, \mathbf{I}]\{1\}$	1157.4	$[\boldsymbol{B}_k, \omega, \mathbf{I}]\{1\}$	1158.4	$[\boldsymbol{B}, \omega_k, \mathbf{I}]\{1\}$	1157.4
	2	$[\boldsymbol{B}, \omega, \boldsymbol{\Lambda}_k]\{1\}$	1160.3	$[\boldsymbol{B}_k, \omega_k, \mathbf{I}]\{1\}$	1161.2	$[\boldsymbol{B}, \omega, \boldsymbol{\Lambda}_k]\{1\}$	1160.3
	3	$[\boldsymbol{B}_k, \omega_k, \mathbf{I}]\{1\}$	1161.2	$[\boldsymbol{B}, \omega, \mathbf{I}]\{1\}$	1163.0	$[\boldsymbol{B}_k, \omega_k, \mathbf{I}]\{1\}$	1161.2
	4	$[\boldsymbol{B}, \omega, \mathbf{I}]\{1\}$	1163.0	$[\boldsymbol{B}, \omega, \boldsymbol{\Lambda}]\{1\}$	1165.7	$[\boldsymbol{B}, \omega, \mathbf{I}]\{1\}$	1163.0
RTV	1	$[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$	1158.8	$[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$	1158.8	$[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$	1158.8
	2	$[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$	1161.4	$[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$	1161.4	$[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$	1161.4
	3	$[\boldsymbol{R}, a_k \boldsymbol{T}, \boldsymbol{V}_k]$	1161.7	$[\boldsymbol{R}, a_k \boldsymbol{T}, \boldsymbol{V}_k]$	1161.7	$[\boldsymbol{R}, a_k \boldsymbol{T}, \boldsymbol{V}_k]$	1161.7
	4	$[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$	1163.4	$[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$	1163.4	$[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$	1163.4

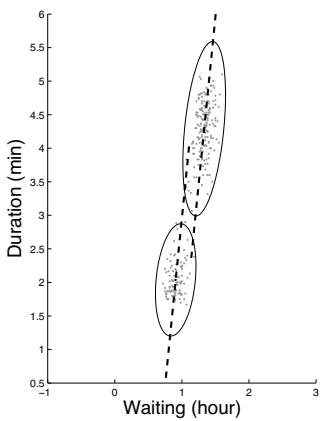
(a) *min*$\times$*min (original units)* (b) *sec*$\times$*min* (c) *standardized*$\times$*standard.*

Graphical units invariance

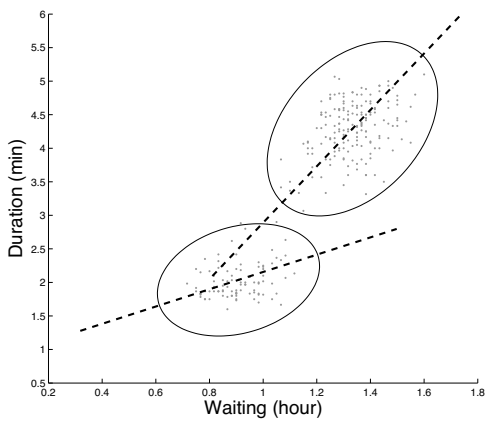
- Graphical representation corresponds to a particular unit choice $\mathbf{u}$

initial unit		*graphical unit*
$\mathbb{X}$	$\xrightarrow{\mathbf{u}}$	$\mathbb{X}^{\mathbf{u}}$
$\mathbf{x}$	$\longmapsto$	$\mathbf{x}^{\mathbf{u}}$
$\mathbf{X} \sim p^{\mathbf{m}}$		$\mathbf{u}(\mathbf{X}) \sim$ "$\mathbf{u}(p^{\mathbf{m}})$" $= p^{\text{"}\mathbf{u}(\mathbf{m})\text{"}} = p^{\mathbf{m}'}$
"standard" model $\mathbf{m}$		"graphical" model "$\mathbf{m}' = \mathbf{u}(\mathbf{m})$"

- Example: EIG models not always invariant to non-isotropic axis rescaling ($\neq$RTV)
  - (a) Gaussians with same orientation in an $\mathbf{u}_1$ =orthonormal basis
  - (b) A modification of $\mathbf{u}_2$=x-axis scale infringes the assumption of same orientations



(a)



(b)

Outline

Partitioning communes of Wallonia

- Data: $n = 262$ communes of Wallonia in terms of $d = 2$ fractals at a local level
  - 1st variable: fractal dimension of city boundary picture
  - 2nd variable: fractal dimension of city surface picture
- See more details in [Thomas *et al.*, 2008][11]

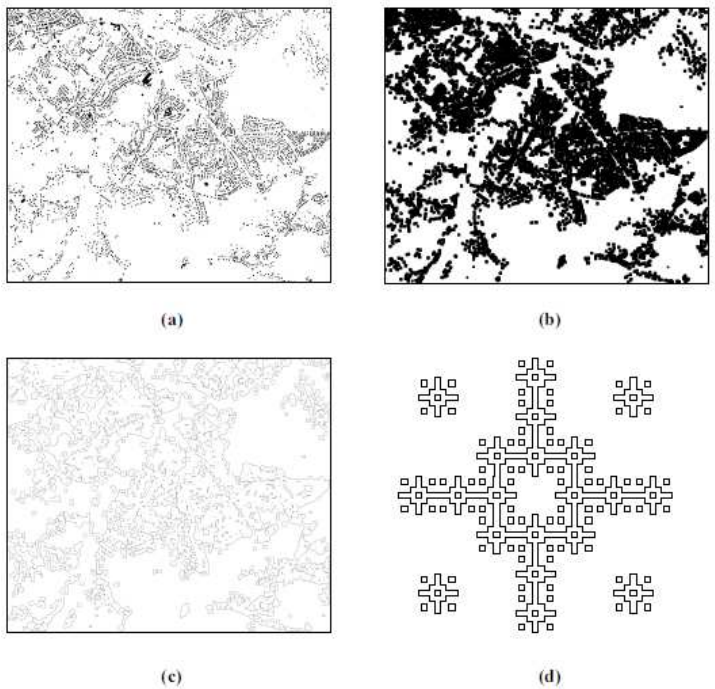


Figure 6: Extracting boundaries by dilation. Figure 6a shows the original urban pattern, Figure 6b the corresponding dilated structure (3 steps), Figure 6c the extracted boundary and Figure 6d a theoretical fractal with similar features to the observed fractal in (c)

[11] I. Thomas, P. Frankhauser and C. Biernacki (2008). The morphology of built-up landscapes in Wallonia (Belgium): a classification using fractal indices. Landscape and Urban Planning, 84, 99-115.

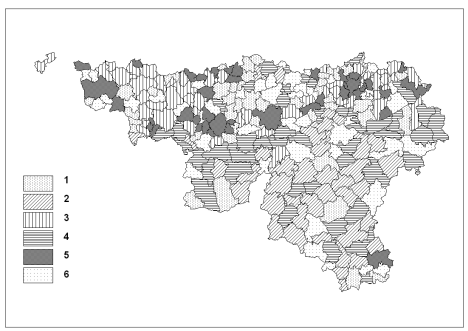
Results for Wallonia

- BIC retains $\mathbf{u} = (\text{exp}, \text{exp})$ and $\mathbf{m} = (\pi_k)[\lambda \mathbf{I}]$ (among **id**/log/exp and all EIG)
- meaningful groups with $\mathbf{u} = (\text{exp}, \text{exp})$
- exp was a natural unit at the fractal level ("fractal dimension")
- exp also natural since it correspond to the "number of pixel pair comparisons"
- Somewhere, exp is quite related to the Manly transformation (see later)

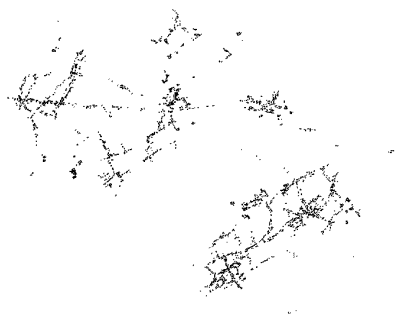
Table 3: Results of the fractal classification of communes

Cluster	$D_{Surf\text{-}d0}$	$D_{bord\text{-}d3}$	n	Three most representative communes	Typology
1	1.37	1.65	44	Brugelette, *Heron*, Nandrin	Peri-urban I and small cities
2	0.92	1.50	40	*Lierneux*, Havelange, Merbes-le-C.	Rural I: compact isolated hamlets
3	1.50	1.76	49	Pepinster, Saint-Georges, *Blegny*	Peri-urban II and eastern part (Hainaut)
4	1.11	1.59	47	Erquelinnes, Baelen, *Rendeux*	Rural II: hamlets with a linear structure
5	1.68	1.70	40	Ottignies, Châtelet, *Chaudfontaine*	Urban (homogeneous, fully urbanised communes)
6	1.25	1.63	42	Gesves, Jalhay, *Ciney*	Rural III: rural communes with hamlets and one (small) city centre

n: number of communes in the class;



Wallonie communes clustering



Heron

Chaudfontaine

Prostate cancer data of [Biar & Green, 1980][14]

- Individuals: 506 patients with prostatic cancer grouped on clinical criteria into two Stages 3 and 4 of the disease
- Variables: $d = 12$ pre-trial variates were measured on each patient, composed by
  - Eight continuous variables (age, weight, systolic blood pressure, diastolic blood pressure, serum haemoglobin, size of primary tumour "SZ", index of tumour stage and histolic grade, serum prostatic acid phosphatase "AP")
  - Two ordinal variables (performance rating, cardiovascular disease history)
  - Two categorical variables with various numbers of levels (electrocardiogram code, bone metastases)
- Some missing data: 62 missing values ($\approx 1\%$)
- Two historical units for performing the clustering task:
  - Raw units **id**: [McParland & Gormley, 2015][12]
  - Transformed data **u**: since SZ and AP are skewed, [Jorgensen & Hunt, 1996][13] propose

$$\mathbf{u}_{SZ} = \sqrt{\cdot} \text{ and } \mathbf{u}_{AP} = \ln(\cdot)$$

[12] McParland, D. and Gormley, I. C. (2015). Model based clustering for mixed data: clustmd. arXiv preprint arXiv:1511.01720.

[13] Jorgensen, M. and Hunt, L. (1996). Mixture model clustering of data sets with categorical and continuous variables. In Proceedings of the Conference ISIS, volume 96, pages 375–384.

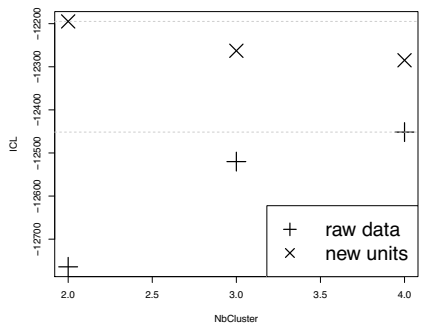
[14] Byar DP, Green SB (1980): Bulletin Cancer, Paris 67:477-488

Clustering with the MixtComp software [Biernacki *et al.*, 2016][15]

- Model **m** in Mixtcomp: full mixed data $\mathbf{x} = (\mathbf{x}^{cont}, \mathbf{x}^{cat}, \mathbf{x}^{ordi}, \mathbf{x}^{int}, \mathbf{x}^{rank})$ (missing data are allowed also) are simply modeled by inter conditional independence

$$p(\mathbf{x}; \boldsymbol{\alpha}_k) = p(\mathbf{x}^{cont}; \boldsymbol{\alpha}_k^{cont}) \times p(\mathbf{x}^{cat}; \boldsymbol{\alpha}_k^{cat}) \times p(\mathbf{x}^{ordi}; \boldsymbol{\alpha}_k^{ordi}) \times \dots$$

- In addition, for symmetry between types, intra conditional independence for each
- Results:
  - New units $\mathbf{u}_{SZ}$ and $\mathbf{u}_{AP}$ are selected by ICL
  - New units allow to select two groups and provides a lower error rate



	clusters	
	1	2
1	287	5
2	52	162

Table : MixtComp model on raw units: 11% misclassified

	clusters	
	1	2
1	270	22
2	23	191

Table : MixtComp model on new units: 9% misclassified

[15] MixtComp is a clustering software developped by Biernacki C., Iovleff I. and Kubicki V. and freely available on the MASSICCC web platform https://modal-research-dev.lille.inria.fr/#/

Outline

Looking for conditional normality

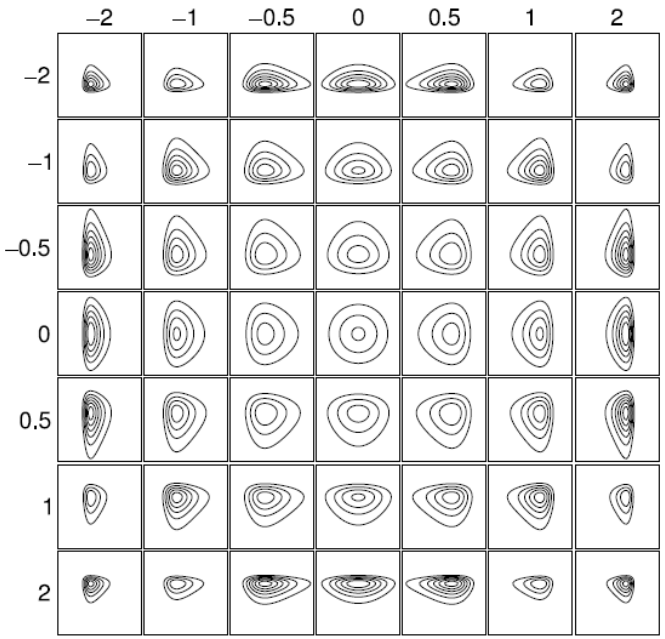
- [Zhu & Melnykov, 2016][16] transform units conditionally to classes for approaching class normality with the Manly transformation unit ($k = 1, \ldots, g, j = 1, \ldots, d$)

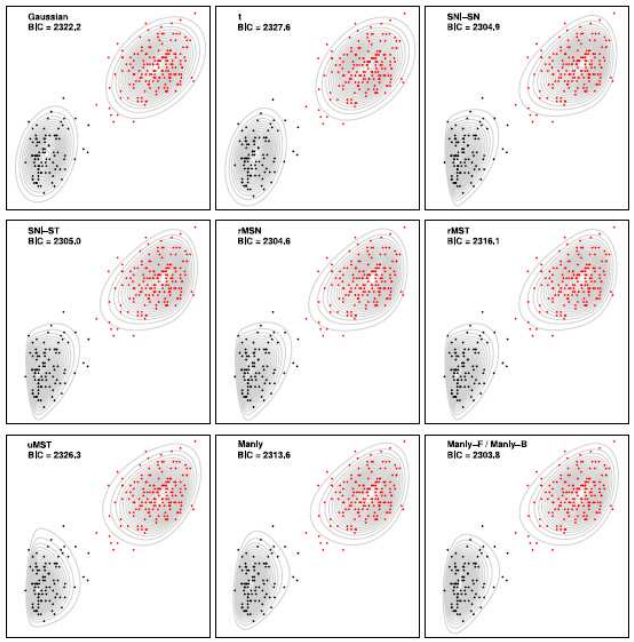
$$\mathbf{u}_{\boldsymbol{\lambda}} = \{\mathbf{u}_{\lambda_{kj}}\} \quad \text{with} \quad \mathbf{u}_{\lambda_{kj}} = \begin{cases} \dfrac{\exp(\lambda_{kj} x^j) - 1}{\lambda_{kj}}, & \lambda_{kj} \neq 0 \\ x^j, & \lambda_{kj} = 0 \end{cases}$$

- Estimate parameters $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ by ml and the EM algorithm
- In fact choosing $\lambda_{kj} \in \{\mathbb{R}^+, \{0\}\}$ corresponds to a model and is performed by a forward and backward selection associated to a BIC criterion

[16] Zhu, X. and Melnykov, V. (2016) Manly Transformation in Finite Mixture Modeling, accepted by Computational Statistics and Data Analysis.

Examples[18]





One bivariate component $\mathcal{N}_2(\mathbf{0}, \mathbf{I})$
Different $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ values

Old Faithful Geyser
[Azzalini & Bowman, 1990][17]

[17] Azzalini, A., Bowman, A.W., 1990. A look at some data on the Old Faithful geyser. J. Roy. Statist. Soc. Ser. C 39, 357–365.

[18] Figures from [Zhu & Melnykov, 2016]

Discussion on Manly units

- High flexibility for mixtures
- But low unit interpretation for two reasons
  - Manly transformation is a non-standard unit (?)
  - Unit transformation is class-dependent...
- Defend invariance of scale transformation of Manly as a desirable property...

2.4. Properties of Manly components

In this section, it is demonstrated that the proposed Manly components are invariant to the shifting and scaling of data points which is extremely desirable in the model-based clustering context as these operations should not lead to a different clustering result. For example, the relationship between temperature measured in degrees Celsius (C) and Fahrenheit (F) is given by $C = 5F/9 - 160/9$ and estimated partitions should be consistent under both scales.

...but it could be an opportunity to have no stability (provide new models!)

Outline

Which units for count data?

- Count data: $x \in \mathbb{N}$
- Standard model $\mathbf{m}$ is Poisson: $\mathrm{p}(\cdot; \alpha_k) = \mathcal{P}(\lambda_k)$
- d-variate case $\mathbf{x} = (x^1, \ldots, x^d) \in \mathbb{N}^d$ and conditional independence by variable
- Two standards unit transformations (by variable $j \in \{1, \ldots, d\}$):
  - Shifted observations: $\mathbf{u}(x^j) = x^j - a_j$ with $a_j \in \mathbb{N}$
  - Scaled observations: $\mathbf{u}(x^j) = b_j x^j$ with $b_j \in \mathbb{N}^*$

Shifted example

- **id**: total number of educational years
- $\mathbf{u}_{shift}(\cdot) = (\cdot) - 8$: university number of educational years[a]

[a] Eight is the number of years spent by english pupils in a secondary school.

Scaled example

- **id**: total number of educational years
- $\mathbf{u}_{scaled}(\cdot) = 2 \times (\cdot)$: total number of educational semesters

Medical data

- R dataset `rwm1984COUNT` of [Rao *et al.*, 2007, p.221][19] and studied in [Hilbe, 2014][20]
- $n = 3874$ patients that spent time into German hospitals during year 1984
- Patients are described through eleven mixed variables
- **m**: a MixtComp model combining Gaussian, Poisson and multinomial distributions

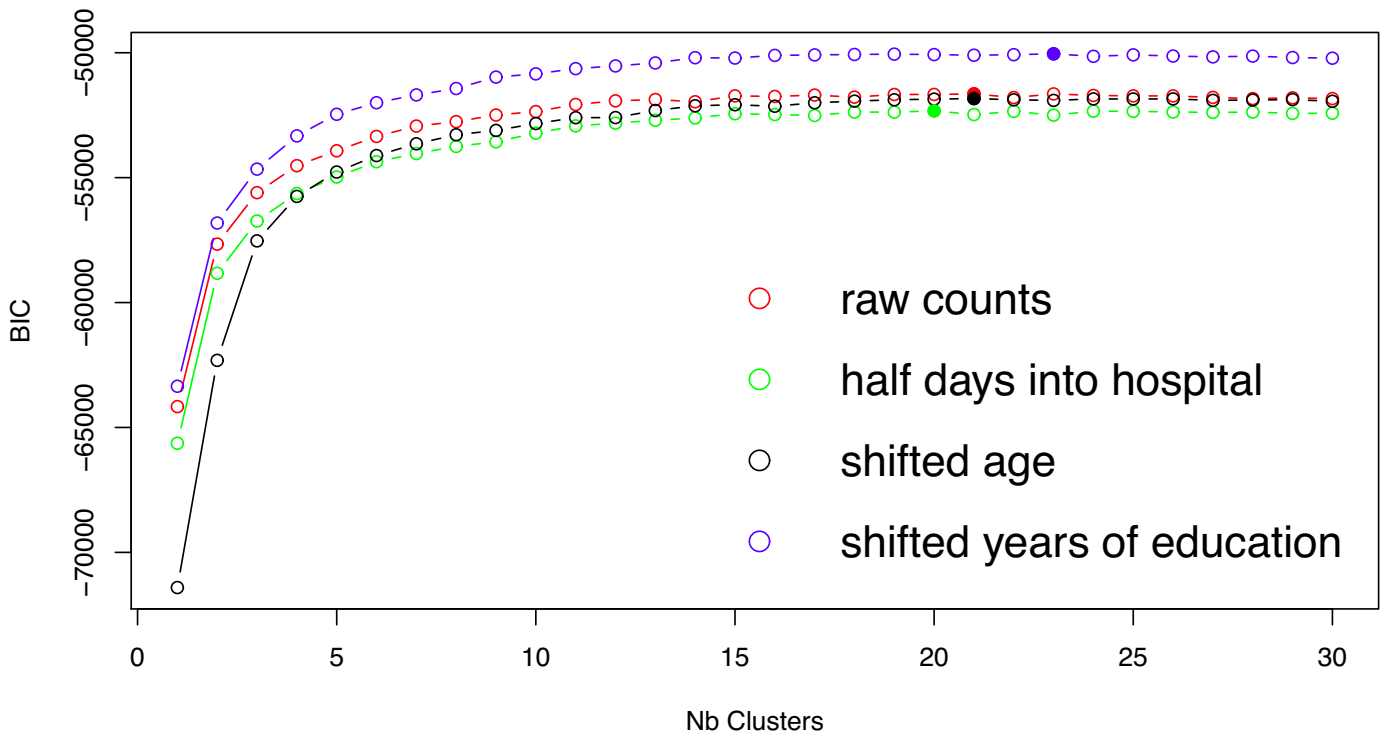
	variables	*type*	*model*
1	number of visits to doctor during year	count	Poisson
2	number of days in hospital	count	Poisson
3	educational level	categorical	multinomial
4	age	count	Poisson
5	outwork	binary	Bernoulli
6	gender	binary	Bernoulli
7	matrimonial status	binary	Bernoulli
8	kids	binary	Bernoulli
9	household yearly income	continous	Gaussian
10	years of education	count	Poisson
11	self employed	binary	Bernoulli

[19] Rao, C. R., Miller, J. P., and Rao, D. C. (2007). Handbook of statistics: epidemiology and medical statistics, volume 27. Elsevier.

[20] Hilbe, J. M. (2014). Modeling count data. Cambridge University Press.

Several units for count data

- Four unit systems are sequentially considered differing over the count data
  - $\mathbf{u}_1 = \mathbf{id}$: original unit
  - $\mathbf{u}_2$: the time spent into hospital is counted in half days instead of days
  - $\mathbf{u}_3$: the minimum of the age series is deduced from all ages leading to shifted ages
  - $\mathbf{u}_4$: the min. of years of edu. is deduced from the series leading to shifted years of edu.
- BIC selects 23 clusters obtained under shifted years of education



Specific transformation for RNA-seq data

- A sample of RNA-seq gene expressions arising from the rat count table of `http://bowtie-bio.sourceforge.net/recount/`
- 30000 genes described by 22 counting descriptors
- Remove genes with low expression (classical): 6173 genes finally
- Two different processes for dealing with data:
  - Standard [Rau *et al.*, 2015][21]: $\mathbf{u} = \mathbf{id}$ and $\mathbf{m}$ is Poisson mixture
  - "RNA-seq unit" [Gallopin *et al.*, 2015][22]:

$$\mathbf{u}(\cdot) = \ln(\text{scaled normalization}(\cdot))$$

    is a transformation being motivated by genetic considerations and $\mathbf{m}$ is Gaussian mixture
  - Experiment with 30 clusters (as in [Gallopin *et al.*, 2015])

model	*data*	*BIC*
Poisson	raw unit	−2615654
Gaussian	transformed	−909190

[21] Rau, A., Maugis-Rabusseau, C. , Martin-Magniette, M.-L. and Celeux, G. (2015). Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. Bioinformatics, 31 (9), 1420-1427.

[22] Gallopin, M., Rau, A., Celeux, G., and Jaffrézic, F. (2015). Transformation des données et comparaison de modèles pour la classification des données rna-seq. In 47èmes Journées de Statistique de la SFdS.

Outline

Co-clustering framework

- It corresponds to the following specific mixture model **m** [Govaert and Nadif, 2014][23]:

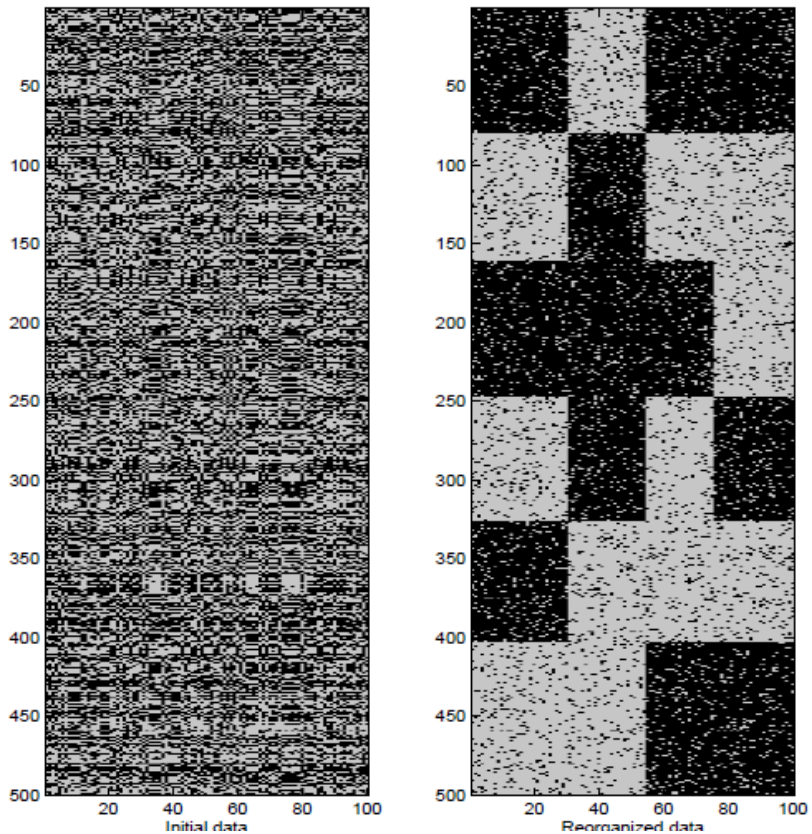
$$\mathrm{p}(\mathbf{x};\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w})} \prod_{i,j} \pi_{z_i} \rho_{w_j} \mathrm{p}(x_i^j;\boldsymbol{\alpha}_{z_i w_j})$$

  - $\mathbf{z}$: partition in g_r rows
  - $\mathbf{w}$: partition in g_c columns
  - $\mathbf{z} \perp \mathbf{w}$ and $x_i^j | (z_i, w_j) \perp x_{i'}^{j'} | (z_{i'}, w_{j'})$
- Distribution $\mathrm{p}(\cdot;\boldsymbol{\alpha}_{z_i w_j})$ depends on the kind of data
  - Binary data: $x_i^j \in \{0,1\}$, $\mathrm{p}(\cdot;\boldsymbol{\alpha}_{kl}) = \mathcal{B}(\alpha_{kl})$
  - Categorical data with m levels: $\mathbf{x}_i^j = \{x_i^{jh}\} \in \{0,1\}^m$ with $\sum_{h=1}^m x_i^{jh} = 1$ and $\mathrm{p}(\cdot;\boldsymbol{\alpha}_{kl}) = \mathcal{M}(\boldsymbol{\alpha}_{kl})$ with $\boldsymbol{\alpha}_{kl} = \{\alpha_k^{jh}\}$
  - Count data: $x_i^j \in \mathbb{N}$, $\mathrm{p}(\cdot;\boldsymbol{\alpha}_{kl}) = \mathcal{P}(\mu_k \nu_l \gamma_{kl})$
  - Continuous data: $x_i^j \in \mathbb{R}$, $\mathrm{p}(\cdot;\boldsymbol{\alpha}_{kl}) = \mathcal{N}(\mu_{kl}, \sigma_{kl}^2)$
- `BlockCluster` [Bhatia *et al.*, 2015][24] is an R package for co-clustering

[23] G. Govaert and M. Nadif (2014). Co-clustering: models, algorithms and applications. ISTE, Wiley. ISBN 978-1-84821-473-6.

[24] P. Bhatia, S. Iovleff, G. Govaert (2015). Blockcluster: An R Package for Model Based Co-Clustering. *Journal of Statistical Software*, in press.

Binary illustration



Outline

SPAM E-mail Database[26]

- $n = 4601$ e-mails composed by 1813 "spams" and 2788 "good e-mails"
- $d = 48 + 6 = 54$ continuous descriptors[25]
  - 48 percentages that a given word appears in an e-mail ("make", "you'...)
  - 6 percentages that a given char appears in an e-mail (";", "$"...)
- Transformation of continuous descriptors into binary descriptors

$$x_i^j = \begin{cases} 1 & \text{if word/char } j \text{ appears in e-mail } i \\ 0 & \text{otherwise} \end{cases}$$

Two different units considered for variable $j \in \{1, \ldots, 54\}$

- $\mathbf{id}_j$: see the previous coding
- $\mathbf{u}_j(\cdot) = 1 - (\cdot)$: reverse the coding

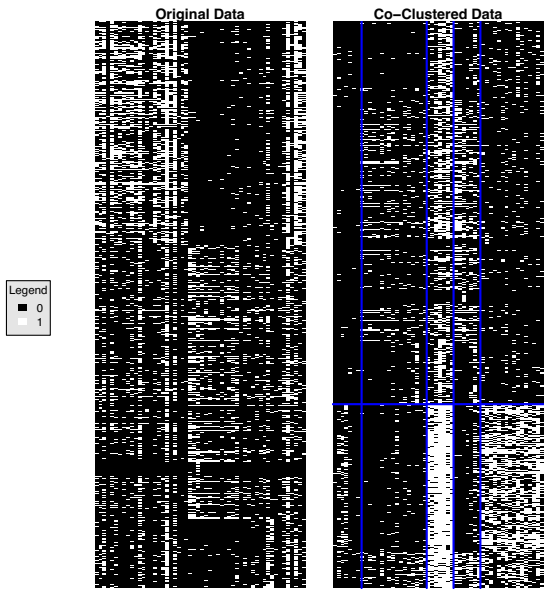
$$\mathbf{u}_j(x_i^j) = \begin{cases} 0 & \text{if word/char } j \text{ appears in e-mail } i \\ 1 & \text{otherwise} \end{cases}$$

[25] There are 3 other continuous descriptors we do not use

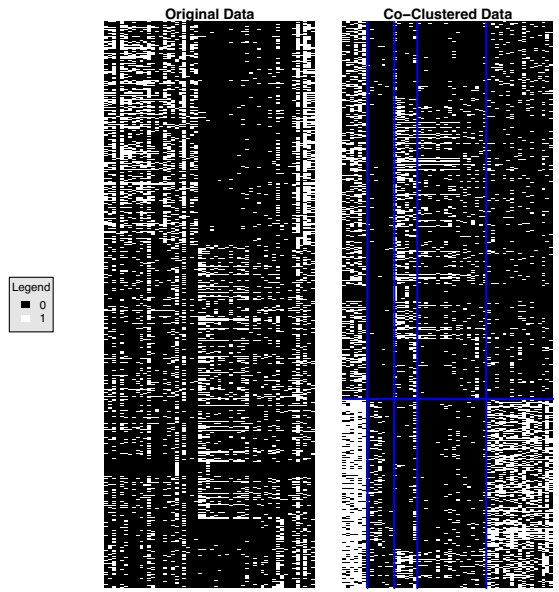
[26] https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/

Select the whole coding $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$

- Fix $g_l = 2$ (two individual classes) and $g_r = 5$ (five variable classes)
- Use co-clustering in a clustering aim: just interested in indiv. classes (spams?)
- Use a "naive" algorithm to find the best $\mathbf{u}$ by ICL (2^{54} possibilities)



initial unit **id**
ICL=-92682.54
error rate=0.1984



best unit **u**
ICL=-92524.57
error rate=0.2008

Result analysis of the e-mail database

- Just one variable ($j = 19$: "you") has a reversed coding in $\mathbf{u}$
- Thus variable "you" has not the same coding as other variables in its column class
- Poor ICL increase with $\mathbf{u}$

Conclusion for the e-mail database

- Here initial units **id** have a particular meaning for the user: do not change!
- In case of unit change, it becomes essentially technic (as Manly unit is)

Outline

Congressional Voting Records Data Set[28]

- Votes for each of the $n = 435$ U.S. House of Representatives Congressmen
- Two classes: 267 democrats, 168 republicans
- $d = 16$ votes with $m = 3$ modalities [Schlimmer, 1987][27]:
  - "yea": voted for, paired for, and announced for
  - "nay": voted against, paired against, and announced against
  - "?": voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known

1. handicapped-infants
2. water-project-cost-sharing
3. adoption-of-the-budget-resolution
4. physician-fee-freeze
5. el-salvador-aid
6. religious-groups-in-schools
7. anti-satellite-test-ban
8. aid-to-nicaraguan-contras
9. mx-missile
10. immigration
11. synfuels-corporation-cutback
12. education-spending
13. superfund-right-to-sue
14. crime
15. duty-free-exports
16. export-administration-act-south-africa

[27] Schlimmer, J. C. (1987). Concept acquisition through representational adjustment. Doctoral dissertation, Department of Information and Computer Science, University of California, Irvine, CA.

[28] http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

Allowed user meaningful recodings

- "yea" and "nea" are arbitrarily coded (question dependent), not "?"
- Example:

  3. adoption-of-the-budget-resolution = "yes" $\Leftrightarrow$ 3. rejection-of-the-budget-resolution = "no"

- However, "?" is not question dependent

Thus, two different units considered for variable $j \in \{1, \ldots, 16\}$

- $\mathbf{id}_j$:

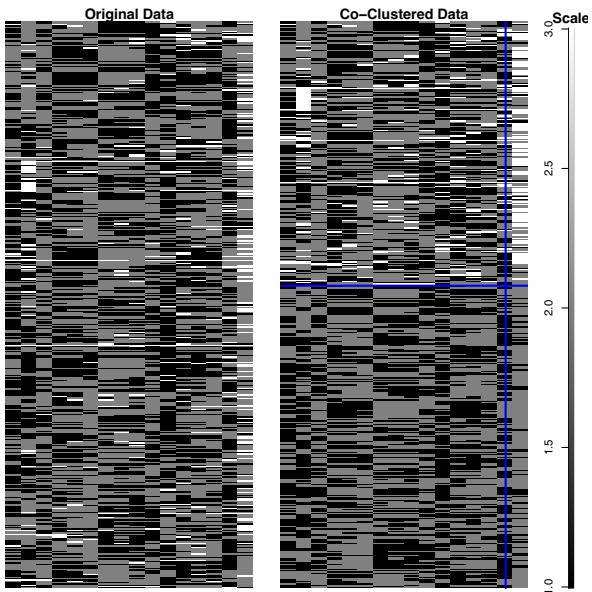
$$x_i^j = \begin{cases} (1,0,0) & \text{if voted "yea" to vote } j \text{ by congressman } i \\ (0,1,0) & \text{if voted "nay" to vote } j \text{ by congressman } i \\ (0,0,1) & \text{if voted "?" to vote } j \text{ by congressman } i \end{cases}$$

- $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$: reverse the coding only for "yea" and "nea"

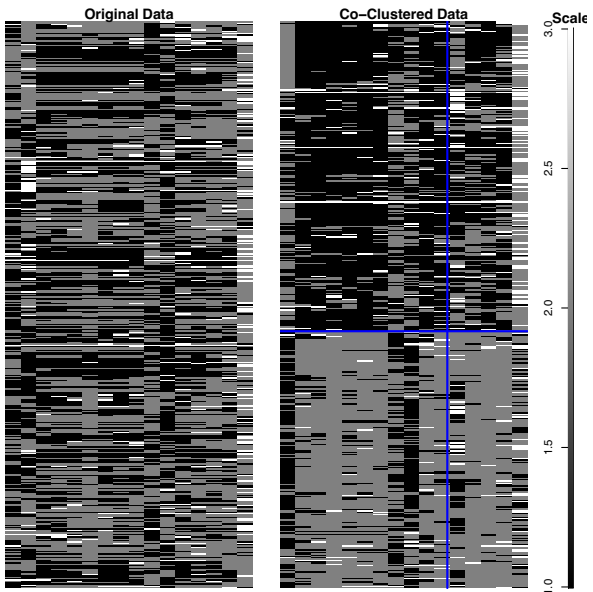
$$\mathbf{u}_j(x_i^j) = \begin{cases} (0,1,0) & \text{if voted "yea" to vote } j \text{ by congressman } i \\ (1,0,0) & \text{if voted "nay" to vote } j \text{ by congressman } i \\ (0,0,1) & \text{if voted "?" to vote } j \text{ by congressman } i \end{cases}$$

Select the whole coding $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$

- Fix $g_l = 2$ (two individual classes) and $g_r = 2$ (two variable classes)
- Use co-clustering in a clustering aim: just interested in political party
- Use a comprehensive algorithm to find the best $\mathbf{u}$ by ICL ($2^{16} = 65536$ cases)



initial unit **id**
ICL=-5916.13
error rate=0.2850



best unit **u**
ICL=-5458.156
error rate=0.1034

Result analysis of the Congressional Voting Records Data Set

- Five variables has a reversed coding in **u**:
  - 3. adoption-of-the-budget-resolution
  - 7. anti-satellite-test-ban
  - 9. aid-to-nicaraguan-contras
  - 10. mx-missile
  - 16. duty-free-exports
- Thus be aware to change the meaning of them when having a look at the figure!
- Significant ICL and error rate improvements with **u**

Conclusion for the Congressional Voting Records

- Here initial units **id** where arbitrary fixed: make sense to change!
- In addition, good improvement...

Outline

Summary

- Be aware that interpretation of ("classical") models is unit dependent
- Models should even be revisited as a couple units $\times$ "classical" models
- Opportunity for cheap/wide/meaningful enlarging of "classical" model families
- But some units could be user meaningful, restricting this "technical enlarging"
- In counterpart, combinatorial problems may occur if the new family is huge

Outline

Units and other data types (and related distributions)

- Ordinal data $x \in \{\text{high grade}, \text{middle grade}, \text{low grade}\}$:
  - **id**: high grade $>$ middle grade $>$ low grade with " $>$"= greater in strength than
  - **u**: low grade $>$ middle grade $>$ high grade with " $>$"= greater in weakness than
  - Related distribution: see [Biernacki & Jacques, 2015][29] and references therein
- Ranking data $x \in \{(\text{car},\text{bike}), (\text{bike},\text{car})\}$:
  - **id**: (car,bike) $\Leftrightarrow$ car is preferred to bike, (bike,car) $\Leftrightarrow$ bike is preferred to car
  - **u**: (car,bike) $\Leftrightarrow$ bike is preferred to car, (bike,car) $\Leftrightarrow$ car is preferred to bike
  - Related distribution: see [Jacques & Biernacki, 2014][30] and references therein
- Other: directional data...

[29] C. Biernacki and J. Jacques (2015). Model-Based Clustering of Multivariate Ordinal Data Relying on a Stochastic Binary Search Algorithm. Statistics and Computing, in press.

[30] J.Jacques & C.Biernacki (2014). Model-based clustering for multivariate partial ranking data. Journal of Statistical and Planning Inference, 149, 201–217.